Biến ngẫu nhiên rời rạc (Discrete Random Variables)

Thống kê máy tính và ứng dụng (ĐTTX)

Vũ Quốc Hoàng (vqhoang@fit.hcmus.edu.vn)
FIT - HCMUS

Ngày 22 tháng 3 năm 2022

Nội dung

1. Biến ngẫu nhiên và phân phối
2. Biến ngẫu nhiên rời rạc và hàm xác suất
3. Các biến ngẫu nhiên độc lập
4. Kì vọng của biến ngẫu nhiên rời rạc
5. Phương sai và độ lệch chuẩn
6. Một số phân phối xác suất rời rạc điển hình
7. Phương pháp xấp xỉ kỳ vọng và phân phối bằng mô phỏng
8. Ví dụ minh họa - Bài toán thu thập phiếu thưởng

Nội dung

1. Biến ngẫu nhiên và phân phối
2. Biến ngẫu nhiên rời rạc và hàm xác suất
3. Các biến ngẫu nhiên độc lập
4. Kì vọng của biến ngẫu nhiên rời rạc
5. Phương sai và độ lệch chuẩn
6. Một số phân phối xác suất rời rạc điển hình
7. Phương pháp xấp xỉ kỳ vọng và phân phối bằng mô phỏng
8. Ví dụ minh họa - Bài toán thu thập phiếu thưởng

Biến ngẫu nhiên

Nếu giá trị của một đại lượng/tính chất X được xác định hoàn toàn khi biết kết quả ω của một thí nghiệm T thì X được gọi là một đại lượng/biến ngẫu nhiên (liên quan đến T)

- Trước khi biết kết quả, ta chỉ biết X có thể nhận một giá trị nào đó trong tập A,
- Sau khi biết kết quả ω, ta biết X nhận một giá trị cụ thể $x \in A$, ta kí hiệu $X(\omega) = x$.

Biến ngẫu nhiên (random variable) là hàm trên không gian mẫu Ω

- $X : \Omega \to A$, gắn mỗi kết quả $\omega \in \Omega$ một giá trị $X(\omega) \in A$,
- A được gọi là **tập/miền giá trị** của X,
- A thường là tập con của tập số thực $\mathbb{R}$.

Biến ngẫu nhiên - Ví dụ

Xét thí nghiệm chọn ngẫu nhiên một sinh viên trong lớp, $\Omega = \{\text{An, Bình, Chương, ...}\}$

- Đo chiều cao H của sinh viên được chọn
  - H là biến số với tập giá trị $\mathbb{R}$ (hoặc tập $[1.0, 2.0]$ mét),
  - $H(\text{An}) = 1.5m, H(\text{Bình}) = 1.7m$, ...
- Xác định giới tính G của sinh viên được chọn
  - G là biến phân loại với tập giá trị $\{\text{Nam, Nữ}\}$ (hoặc tập $\{0, 1\}$),
  - $G(\text{An}) = \text{Nữ}, G(\text{Bình}) = \text{Nam}$, ...
- Xét điểm của sinh viên được chọn: biến số với tập giá trị $\{0, 0.5, 1, 1.5, ..., 9.5, 10\}$ (hoặc tập $\mathbb{R}$).
- Xét học lực của sinh viên được chọn: biến phân loại với tập giá trị {Yếu, Kém, Trung-bình, Khá, Giỏi, Xuất-sắc}.

Biến ngẫu nhiên

Biến ngẫu nhiên là phương tiện hay được dùng để mô tả các biến cố. Xét biến (số) ngẫu nhiên X liên quan đến thí nghiệm T có không gian mẫu là Ω. Cho $C \subset \mathbb{R}$, ta kí hiệu biến cố "X nhận giá trị trong C" là

$$(X \in C) = \{\omega \in \Omega : X(\omega) \in C\}.$$

- Chẳng hạn, cho $x \in \mathbb{R}$ ta kí hiệu

$$(X = x) = \{\omega \in \Omega : X(\omega) = x\},$$
$$(X \leq x) = \{\omega \in \Omega : X(\omega) \leq x\},$$

- Hay với hai biến X, Y ta kí hiệu

$$(X = Y) = \{\omega \in \Omega : X(\omega) = Y(\omega)\},$$
$$(X \leq Y) = \{\omega \in \Omega : X(\omega) \leq Y(\omega)\}.$$

- Các biến cố này còn được gọi là biến cố liên quan đến biến ngẫu nhiên X, Y.

Biến ngẫu nhiên - Ví dụ

Xét thí nghiệm tung một xúc xắc (đồng chất) 2 lần, $\Omega = \{(i,j) : i,j \in \{1,...,6\}\}$, mô hình xác suất đơn giản. Gọi X, Y là số chấm ở lần 1, 2

$$X(\omega = (i,j)) = i \quad \text{và} \quad Y((i,j)) = j.$$

- Biến cố "số chấm ở lần một là 6" là $(X = 6) = \{(6,j) : j \in \{1,...,6\}\}$.
- Biến cố "số chấm ở hai lần giống nhau" là $(X = Y) = \{(i,j) : i = j \in \{1,...,6\}\}$

$$P(X = Y) = \frac{|(X = Y)|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}.$$

- Xác suất được "số chấm ở lần một lớn hơn số chấm ở lần hai" khi biết "số chấm ở lần hai lớn hơn 4" là

$$P(X > Y | Y > 4) = \frac{P(X > Y > 4)}{P(X > 4)} = \frac{|(X > Y > 4)|}{|(X > 4)|} = \frac{1}{12}.$$

Phân phối của biến ngẫu nhiên

Xét biến ngẫu nhiên X liên quan đến thí nghiệm T có không gian mẫu là Ω. Tập các xác suất $\{P(X \in C) : C \subset \mathbb{R}\}$ xác định một độ đo xác suất trên (không gian mẫu mới) $\mathbb{R}$ và được gọi là **phân phối** (distribution) của X.

- Phân phối của X cho thấy khả năng X nhận các giá trị khác nhau.
- Với phân phối của X, ta khảo sát X mà không cần để ý đến T hay Ω nữa.
- Nói chung, tập $\{P(X \in C) : C \subset \mathbb{R}\}$ là "rất khó tính toán". Ta cần cách nào đó giúp xác định phân phối của X để "dễ tính toán hơn".

Phân phối của biến ngẫu nhiên - Ví dụ

Xét biến ngẫu nhiên X có tập giá trị là $\{x_0\}$

- $X(\omega) = x_0$ với mọi $\omega \in \Omega$.
- X chỉ có 2 biến cố liên quan là

$$(X \neq x_0) = \varnothing \quad \text{và} \quad (X = x_0) = \Omega.$$

- "Không nên gọi X là biến ngẫu nhiên" vì ta biết giá trị của X chắc chắn là x_0.
- Phân phối của X rất đơn giản

$$P(X \in C) = \begin{cases} 1 & \text{nếu } x_0 \in C, \\ 0 & \text{nếu } x_0 \notin C. \end{cases}$$

Ví dụ: xét biến ngẫu nhiên X là "điểm tổng kết" trong thí nghiệm "bỏ thi môn TKMT&UD", X chỉ có một giá trị là 0 (điểm).

Nội dung

1. Biến ngẫu nhiên và phân phối
2. Biến ngẫu nhiên rời rạc và hàm xác suất
3. Các biến ngẫu nhiên độc lập
4. Kì vọng của biến ngẫu nhiên rời rạc
5. Phương sai và độ lệch chuẩn
6. Một số phân phối xác suất rời rạc điển hình
7. Phương pháp xấp xỉ kỳ vọng và phân phối bằng mô phỏng
8. Ví dụ minh họa - Bài toán thu thập phiếu thưởng

Biến ngẫu nhiên rời rạc và hàm xác suất

- X được gọi là **biến ngẫu nhiên rời rạc** (discrete random variable) nếu tập giá trị của nó là **rời rạc** (**hữu hạn** (finite) hoặc **vô hạn đếm được** (countably infinite)).
- Với X là biến ngẫu nhiên rời rạc, **hàm xác suất** (probability function, probability mass function) của X là hàm $f : \mathbb{R} \to \mathbb{R}$, được xác định bởi

$$f(x) = f_X(x) = P(X = x), x \in \mathbb{R}.$$

  - Hàm xác suất f cho biết khả năng X nhận một giá trị cụ thể.
  - Tập số thực $\{x \in \mathbb{R} : f(x) > 0\}$ được gọi là tập hỗ trợ (support) của X, kí hiệu $\text{Sup}(X)$.
  - Hàm xác suất có tính chất: $f(x) \geq 0, \forall x \in \mathbb{R}$ và $\sum_{x \in \text{Sup}(X)} f(x) = 1$.
- Hàm xác suất xác định phân phối của biến ngẫu nhiên rời rạc

$$P(X \in C) = \sum_{x \in C} f(x), C \subset \mathbb{R}.$$

Biến ngẫu nhiên rời rạc và hàm xác suất - Ví dụ

Xét thí nghiệm tung một đồng xu (đồng chất) 2 lần, gọi X là số lần được mặt ngửa

- X là biến ngẫu nhiên rời rạc với tập giá trị $\{0, 1, 2\}$,
- X có hàm xác suất

$$f_X(x) = P(X = x) = \begin{cases} 1/4 & \text{nếu } x = 0, \\ 2/4 & \text{nếu } x = 1, \\ 1/4 & \text{nếu } x = 2, \\ 0 & \text{nếu } x \notin \{0, 1, 2\}. \end{cases}$$

- Hàm f_X còn được cho bởi bảng sau, gọi là **bảng phân phối xác suất** của X

x	0	1	2
$f_X(x)$	1/4	1/2	1/4

Hàm đặc trưng của biến cố

Cho biến cố E liên quan đến thí nghiệm T với không gian mẫu Ω, ta gọi **hàm đặc trưng** (characteristic function, indicator function) của E là hàm $\mathbb{I}_E : \Omega \to \mathbb{R}$ được xác định bởi

$$\mathbb{I}_E(\omega) = \begin{cases} 1 & \text{nếu } \omega \in E, \\ 0 & \text{nếu } \omega \notin E. \end{cases}$$

Hàm đặc trưng giúp khảo sát biến cố như là một biến ngẫu nhiên.

Ví dụ: xét thí nghiệm tung một xúc xắc đồng chất, gọi E là biến cố "được mặt lớn hơn 4" thì $\mathbb{I}_E$ có thể được xem là "số lần được mặt lớn hơn 4". $\mathbb{I}_E$ là biến ngẫu nhiên rời rạc với bảng phân phối xác suất

x	0	1
$f_{\mathbb{I}_E}(x)$	2/3	1/3

Phân phối đều rời rạc

Biến ngẫu nhiên X được gọi là có **phân phối đều rời rạc** (discrete uniform distribution) trên tập n giá trị $A = \{x_1, x_2, ..., x_n\}$ nếu X có hàm xác suất

$$f_X(x) = P(X = x) = \frac{1}{n}, x \in A.$$

X là kết quả của thí nghiệm "chọn ngẫu nhiên một giá trị trong tập A".

Ví dụ: xét thí nghiệm tung một xúc xắc đồng chất 2 lần, gọi X, Y là số chấm ở lần 1, 2

- X, Y là các biến ngẫu nhiên rời rạc có phân phối đều trên tập $\{1, 2, ..., 6\}$.
- Tuy nhiên, "tổng số chấm của hai lần", $Z = X + Y$ là biến ngẫu nhiên rời rạc với tập giá trị $\{2, 3, ..., 11, 12\}$ có phân phối không đều.

Nội dung

1. Biến ngẫu nhiên và phân phối
2. Biến ngẫu nhiên rời rạc và hàm xác suất
3. Các biến ngẫu nhiên độc lập
4. Kì vọng của biến ngẫu nhiên rời rạc
5. Phương sai và độ lệch chuẩn
6. Một số phân phối xác suất rời rạc điển hình
7. Phương pháp xấp xỉ kỳ vọng và phân phối bằng mô phỏng
8. Ví dụ minh họa - Bài toán thu thập phiếu thưởng

Các biến ngẫu nhiên độc lập

Hai biến ngẫu nhiên X, Y được gọi là **độc lập** (independent) nếu với mọi $A, B \subset \mathbb{R}$ ta có

$$P\left((X \in A) \cap (Y \in B)\right) = P(X \in A)P(Y \in B).$$

Nghĩa là việc X nhận giá trị nào cũng không ảnh hưởng đến khả năng nhận giá trị nào đó của Y (và ngược lại).

Mệnh đề. Hai biến ngẫu nhiên rời rạc X, Y độc lập khi và chỉ khi

$$P\left((X = x) \cap (Y = y)\right) = P(X = x)P(Y = y)$$

với mọi $x, y \in \mathbb{R}$.

Ta cũng kí hiệu $P(X = x, Y = y)$ để chỉ $P((X = x) \cap (Y = y))$.

Các biến ngẫu nhiên độc lập - Ví dụ

Xét thí nghiệm tung một xúc xắc đồng chất 2 lần, gọi X, Y là số chấm ở lần 1, 2 và $S = X + Y$ là tổng số chấm của 2 lần.

- X, Y độc lập vì

$$P(X = x, Y = y) = \begin{cases} \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = P(X = x)P(Y = y) & x, y \in \{1, 2, ..., 6\} \\ 0 = P(X = x)P(y = y) & \text{khác.} \end{cases}$$

- X, S không độc lập vì

$$P(X = 2) = \frac{1}{6} \quad \text{và} \quad P(S = 2) = P(X + Y = 2) = P(X = 1, Y = 1) = \frac{1}{36}$$

nhưng $P(X = 2, S = 2) = 0 \neq P(X = 2)P(S = 2)$.
Ta cũng có

$$P(S = 2|X = 2) = 0 \neq P(S = 2) = \frac{1}{36}.$$

Nội dung

1. Biến ngẫu nhiên và phân phối
2. Biến ngẫu nhiên rời rạc và hàm xác suất
3. Các biến ngẫu nhiên độc lập
4. Kì vọng của biến ngẫu nhiên rời rạc
5. Phương sai và độ lệch chuẩn
6. Một số phân phối xác suất rời rạc điển hình
7. Phương pháp xấp xỉ kỳ vọng và phân phối bằng mô phỏng
8. Ví dụ minh họa - Bài toán thu thập phiếu thưởng

Kì vọng của biến ngẫu nhiên rời rạc

Cho biến ngẫu nhiên rời rạc X với hàm xác suất f, **kì vọng** (mean, expectation, average) của X, kí hiệu $E(X)$, là số thực được tính bởi ("nếu tính được")

$$E(X) = \mu_X = \mu = \sum_x xP(X = x) = \sum_x xf(x).$$

Kì vọng của X là giá trị trung bình của các giá trị mà X có thể nhận với trọng số là xác suất để X nhận các giá trị tương ứng đó.

Ví dụ: cho biến cố E với $P(E) = p$, khi đó $\mathbb{I}_E$ có kì vọng

$$\begin{aligned} E(\mathbb{I}_E) &= \sum_x xP(X = x) = \sum_{x \in \{0,1\}} xP(X = x) \\ &= 0 \times P(X = 0) + 1 \times P(X = 1) = 0 \times (1 - p) + 1 \times p \\ &= p = P(E). \end{aligned}$$

Kì vọng của biến ngẫu nhiên rời rạc - Ví dụ 1

Một lớp học gồm 20 SV có "phổ điểm" môn Xác suất như sau

Điểm	4	6	7	8	9	10
Số SV	4	5	5	3	2	1

Chọn ngẫu nhiên một SV trong lớp, gọi X là điểm của SV thì X là biến ngẫu nhiên rời rạc có bảng phân phối xác suất

x	4	6	7	8	9	10
$P(X=x)$	$4/20$	$5/20$	$5/20$	$3/20$	$2/20$	$1/20$

Kì vọng của X là

$$E(X) = 4 \times \frac{4}{20} + 6 \times \frac{5}{20} + 7 \times \frac{5}{20} + 8 \times \frac{3}{20} + 9 \times \frac{2}{20} + 10 \times \frac{1}{20} = 6.65$$

Nhận xét: $E(X)$ chính là điểm trung bình của lớp.

Kì vọng của biến ngẫu nhiên rời rạc - Ví dụ 2

Xét thí nghiệm tung một đồng xu đồng chất 2 lần, gọi X là số lần được mặt ngửa. X là biến ngẫu nhiên rời rạc có bảng phân phối xác suất

x	0	1	2
$P(X = x)$	$1/4$	$1/2$	$1/4$

Ta có kì vọng của X là

$$E(X) = \sum_{x \in \{0,1,2\}} xP(X = x) = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$$

Vậy, 2 lần tung thì "kì vọng" (trung bình) được 1 lần ngửa.

Hàm của biến ngẫu nhiên và kì vọng

Cho biến ngẫu nhiên $X : \Omega \to \mathbb{R}$ và hàm số $r : \mathbb{R} \to \mathbb{R}$, khi đó hàm hợp $Y = r \circ X : \Omega \to \mathbb{R}$ được cho bởi

$$Y(\omega) = r(X(\omega)), \omega \in \Omega$$

cũng là biến ngẫu nhiên, kí hiệu $Y = r(X)$.

Mệnh đề. Nếu X là biến ngẫu nhiên rời rạc và $Y = r(X)$ thì Y là biến ngẫu nhiên rời rạc với kỳ vọng

$$E(Y) = E(r(X)) = \sum_x r(x)P(X = x).$$

Hàm của biến ngẫu nhiên và kì vọng - Ví dụ

Nhà cái rung hai đồng xu (đồng chất). Người chơi sẽ được 1$ nếu không ra ngửa, mất 1$ nếu ra hai ngửa và không được/mất gì nếu ra một ngửa. Gọi X là số mặt ngửa, Y là số tiền người chơi kiếm được thì $Y = r(X)$ với hàm số r được xác định bởi

$$r(x) = \begin{cases} 1 & x = 0, \\ 0 & x = 1, \\ -1 & x = 2. \end{cases}$$

Ta có bảng phân phối xác suất

x	0	1	2
$y = r(x)$	1	0	-1
$P(X = x)$	1/4	1/2	1/4

Kì vọng của Y là $E(Y) = \sum_x r(x)P(X = x) = 1 \times \frac{1}{4} + 0 \times \frac{1}{2} + (-1) \times \frac{1}{4} = 0\$$.

Nội dung

1. Biến ngẫu nhiên và phân phối
2. Biến ngẫu nhiên rời rạc và hàm xác suất
3. Các biến ngẫu nhiên độc lập
4. Kì vọng của biến ngẫu nhiên rời rạc
5. **Phương sai và độ lệch chuẩn**
6. Một số phân phối xác suất rời rạc điển hình
7. Phương pháp xấp xỉ kỳ vọng và phân phối bằng mô phỏng
8. Ví dụ minh họa - Bài toán thu thập phiếu thưởng

Phương sai và độ lệch chuẩn

Cho biến ngẫu nhiên rời rạc X với hàm xác suất f và kì vọng $\mu = E(X)$, **phương sai** (variance) của X, kí hiệu $Var(X)$, là số thực được tính bởi ("nếu tính được")

$$Var(X) = \sigma_X^2 = \sigma^2 = E\left((X-\mu)^2\right) = \sum_x (x-\mu)^2 P(X=x) = \sum_x (x-\mu)^2 f(x).$$

Khi đó ta cũng nói $\sigma = \sigma_X = \sqrt{\sigma_X^2} = \sqrt{Var(X)}$ là **độ lệch chuẩn** (standard deviation) của X. *Lưu ý:* độ lệch chuẩn có cùng đơn vị với X nhưng phương sai thì không.

Phương sai (và độ lệch chuẩn) phản ánh sự **phân tán** của phân phối của biến ngẫu nhiên.

Mệnh đề. Cho X là biến ngẫu nhiên (có phương sai), ta có

$$Var(X) = E(X^2) - (E(X))^2.$$

Phương sai và độ lệch chuẩn - Ví dụ 1

Một lớp học gồm 20 SV có "phổ điểm" các môn Toán, Lý, Hóa như sau

Điểm	1	2	3	4	5	6	7	8	9	10
Toán	0	0	0	0	10	10	0	0	0	0
Lý	10	0	0	0	0	0	0	0	0	10
Hóa	2	2	2	2	2	2	2	2	2	2

Chọn ngẫu nhiên một SV trong lớp, gọi X_1, X_2, X_3 lần lượt là điểm môn Toán, Lý, Hóa của SV. Tính toán ta có

$$\mu_1 = \mu_2 = \mu_3 = 5.5,$$

$$\sigma_1 = 0.50, \sigma_2 = 4.50, \sigma_3 = 2.87.$$

Nhận xét: 3 môn có điểm trung bình như nhau nhưng phổ điểm môn Lý rộng nhất và phổ điểm môn Toán hẹp nhất. Như thông thường, ta mô tả điểm Toán của lớp là 5.5 ± 0.50, điểm Lý là 5.5 ± 4.50, điểm Hóa là 5.5 ± 2.87.

Phương sai và độ lệch chuẩn - Ví dụ 2

Cho biến cố E với $P(E) = p$, khi đó $\mathbb{I}_E$ có kì vọng $\mu = E(\mathbb{I}_E) = p$ và phương sai

$$\begin{aligned} Var(\mathbb{I}_E) &= E\left((\mathbb{I}_E - \mu)^2\right) = \sum_x (x-\mu)^2 P(X=x) = \sum_{x\in\{0,1\}} (x-\mu)^2 P(X=x) \\ &= (0-p)^2 \times P(X=0) + (1-p)^2 \times P(X=1) \\ &= p^2(1-p) + (1-p)^2 p = p(1-p)(p+1-p) \\ &= p(1-p). \end{aligned}$$

Ta cũng có

$$E(\mathbb{I}_E^2) = \sum_{x\in\{0,1\}} x^2 P(X=x) = 0^2 \times (1-p) + 1^2 \times p = p,$$

nên

$$Var(\mathbb{I}_E) = E(\mathbb{I}_E^2) - (E(\mathbb{I}_E))^2 = p - p^2 = p(1-p).$$

Các tính chất quan trọng của kì vọng và phương sai

Cho $X_1, X_2, ..., X_n$ là các biến ngẫu nhiên (có kỳ vọng), ta có

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) \quad (\textbf{linearity of expectation})$$

Cho X là biến ngẫu nhiên và a, b là các hằng số thực, ta có

1. $E(aX + b) = aE(X) + b$,
2. $Var(aX + b) = a^2 Var(X)$.

Cho X, Y là hai biến ngẫu nhiên độc lập, ta có

1. $E(XY) = E(X)E(Y)$,
2. $Var(X + Y) = Var(X) + Var(Y)$.

Nội dung

1. Biến ngẫu nhiên và phân phối
2. Biến ngẫu nhiên rời rạc và hàm xác suất
3. Các biến ngẫu nhiên độc lập
4. Kì vọng của biến ngẫu nhiên rời rạc
5. Phương sai và độ lệch chuẩn
6. **Một số phân phối xác suất rời rạc điển hình**
7. Phương pháp xấp xỉ kỳ vọng và phân phối bằng mô phỏng
8. Ví dụ minh họa - Bài toán thu thập phiếu thưởng

Phân phối Bernoulli

Biến ngẫu nhiên rời rạc X được gọi là có **phân phối Bernoulli** (Bernoulli distribution) với tham số p $(0 \leq p \leq 1)$, kí hiệu $X \sim \text{Bernoulli}(p)$, nếu X có tập giá trị là $\{0, 1\}$ và

$$f(x) = P(X = x) = \begin{cases} p & \text{nếu } x = 1, \\ 1 - p & \text{nếu } x = 0. \end{cases}$$

Khi đó, X có kì vọng $E(X) = p$ và phương sai $Var(X) = p(1 - p)$.

Xét thí nghiệm tung một đồng xu với xác suất ra ngửa p, gọi X là "số lần được ngửa" thì $X \sim \text{Bernoulli}(p)$. Trường hợp đồng xu đồng chất thì $X \sim \text{Bernoulli}(0.5)$.

Xét thí nghiệm T với biến cố A có $P(A) = p$, khi đó $\mathbb{I}_A \sim \text{Bernoulli}(p)$.

Phân phối nhị thức

Biến ngẫu nhiên rời rạc X được gọi là có **phân phối nhị thức** (binomial distribution) với tham số n ($n \in \mathbb{N}$), p ($0 \leq p \leq 1$), kí hiệu $X \sim \mathcal{B}(n, p)$, nếu X có tập giá trị là $\{0, 1, ..., n\}$ và

$$f(x) = P(X = x) = C_n^x p^x (1-p)^{n-x}, x \in \{0, 1, ..., n\}.$$

Khi đó, X có kì vọng $E(X) = np$ và phương sai $Var(X) = np(1-p)$.

Cho thí nghiệm T với biến cố A có $P(A) = p$. Xét thí nghiệm R "thực hiện T lặp lại n lần độc lập", gọi X là "số lần A xảy ra" thì $X \sim \mathcal{B}(n, p)$.

Mệnh đề. Nếu $X_1, X_2, ..., X_n$ là các biến ngẫu nhiên **độc lập và cùng phân phối** (independent and identically distributed - iid) Bernoulli với tham số p, thường kí hiệu $X_1, X_2, ..., X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, và $X = \sum_{i=1}^n X_i$ thì $X \sim \mathcal{B}(n, p)$.

Phân phối nhị thức - Ví dụ

Đề thi gồm 50 câu trắc nghiệm, mỗi câu trắc nghiệm chọn một trong 4 lựa chọn. Chọn đáp án ngẫu nhiên cho mỗi câu, gọi X là số câu đúng thì $X \sim \mathcal{B}(n = 50, p = 1/4)$.

- Xác suất được 5/10 điểm: $P(X = 25) = C_{50}^{25}(1/4)^{25}(3/4)^{25} = 8.45 \times 10^{-5}$.
- Xác suất được không quá 2 điểm

$$P(X \leq 10) = \sum_{x=0}^{10} C_{50}^{x}(1/4)^{x}(3/4)^{50-x} = 0.26.$$

- Xác suất được từ 8 điểm trở lên

$$P(X \geq 40) = \sum_{x=40}^{50} C_{50}^{x}(1/4)^{x}(3/4)^{50-x} = 5.2 \times 10^{-16}.$$

- Kì vọng điểm đạt được: $E(\frac{10}{50}X) = \frac{10}{50}E(X) = \frac{10}{50} \times 50 \times \frac{1}{4} = 2.5$.

Phân phối hình học

Biến ngẫu nhiên rời rạc X được gọi là có **phân phối hình học** (geometric distribution) với tham số p $(0 < p \leq 1)$, kí hiệu $X \sim \text{Geometric}(p)$, nếu X có tập giá trị là $\{1, 2, ...\}$ và

$$f(x) = P(X = x) = (1 - p)^{x-1}p, x \in \{1, 2, ...\}.$$

Khi đó, X có kì vọng $E(X) = \frac{1}{p}$ và phương sai $Var(X) = \frac{1-p}{p^2}$.

Cho thí nghiệm T với biến cố A có $P(A) = p$. Xét thí nghiệm R "thực hiện T lặp lại nhiều lần độc lập cho đến khi A xảy ra thì dừng", gọi X là "số lần thực hiện" thì $X \sim \text{Geometric}(p)$.

Mệnh đề (tính không nhớ - memoryless). Cho $X \sim \text{Geometric}(p)$, với mọi $n = 1, 2, ...$ và mọi $k = 0, 1, ...$ ta có

$$P(X = k + n|X > k) = P(X = n).$$

Phân phối hình học - Ví dụ

Xét thí nghiệm bắn đạn vào bia cho đến khi trúng thì dừng. Giả sử các lần bắn độc lập nhau với cùng xác suất trúng là 1/20, gọi X là số viên đạn phải dùng thì $X \sim \text{Geometric}(p = 0.05)$.

- Xác suất dùng đúng 6 viên đạn: $P(X = 6) = (1 - 0.05)^5 0.05 = 0.0387$.
- Xác suất dùng không quá 5 viên đạn

$$P(X \leq 5) = \sum_{x=1}^{5} (1 - 0.05)^{x-1} 0.05 = 0.2262.$$

- Xác suất dùng nhiều hơn 100 viên đạn

$$P(X > 100) = 1 - P(X \leq 100) = 1 - \sum_{x=1}^{100} (1 - 0.05)^{x-1} 0.05 = 0.0059.$$

- Kì vọng số viên đạn phải dùng: $E(X) = \frac{1}{p} = \frac{1}{0.05} = 20$.

Phân phối Poisson

Biến ngẫu nhiên rời rạc X được gọi là có **phân phối Poisson** (Poisson distribution) với tham số λ ($\lambda > 0$), kí hiệu $X \sim \text{Poisson}(\lambda)$, nếu X có tập giá trị là $\{0, 1, 2, ...\}$ và

$$f(x) = P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, x \in \{0, 1, 2, ...\}.$$

Khi đó, X có kì vọng $E(X) = \lambda$ và phương sai $Var(X) = \lambda$.

Mệnh đề. Cho $X_1 \sim \text{Poisson}(\lambda_1), X_2 \sim \text{Poisson}(\lambda_2)$ và X_1, X_2 độc lập, ta có $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Mệnh đề. Cho $X \sim \mathcal{B}(n, p = \frac{\lambda}{n})$ với hằng số $\lambda > 0$, ta có

$$\lim_{n\to\infty} f_X(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \text{ với mọi } x \in \{0, 1, 2, ...\}.$$

Khi n lớn và p nhỏ, ta có thể xấp xỉ phân phối $\mathcal{B}(n, p)$ bằng phân phối $\text{Poisson}(\lambda)$.

Phân phối Poisson - Ví dụ

Một cuốn sách được xuất bản dày 250 trang. Tỉ lệ lỗi đánh máy loại A là 2 lỗi và loại B là 1.5 lỗi trên 100 trang. Giả sử 2 loại lỗi này xuất hiện độc lập.

- Xác suất cuốn sách không có lỗi loại A? Gọi X là số lỗi loại A trong sách, ta mô hình X bằng phân phối Poisson với kì vọng $\lambda_1 = 2 \times \frac{250}{100} = 5$, khi đó

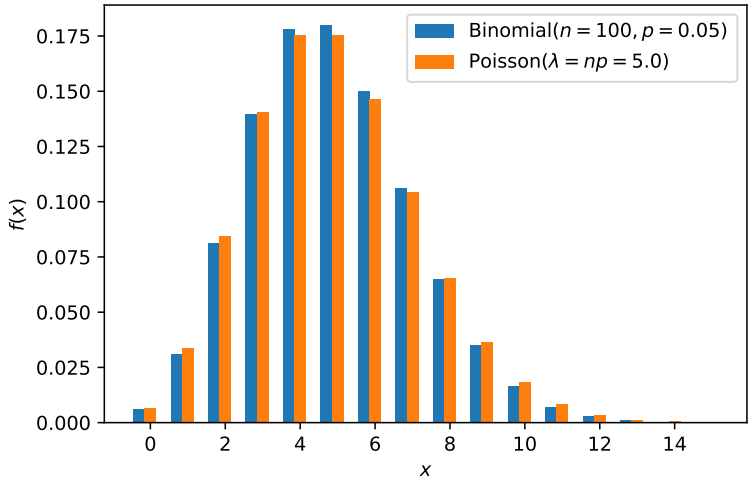
$$P(X = 0) = e^{-\lambda_1} \frac{\lambda_1^0}{0!} = e^{-5} \frac{1}{1} = 0.0067.$$

- Xác suất cuốn sách có ít nhất 5 lỗi loại B? Gọi Y là số lỗi loại B trong sách, ta mô hình Y bằng phân phối Poisson với kì vọng $\lambda_2 = 1.5 \times \frac{250}{100} = 3.75$, khi đó

$$P(Y \geq 5) = 1 - P(Y < 5) = 1 - \sum_{y=0}^{4} e^{-\lambda_2} \frac{\lambda_2^y}{y!} = 1 - \sum_{y=0}^{4} e^{-3.75} \frac{3.75^y}{y!} = 0.3225.$$

- Kì vọng số lỗi cả 2 loại xuất hiện trong sách? Đặt $S = X + Y$ thì $S \sim \text{Poisson}(\lambda_1 + \lambda_2)$, do đó $E(S) = \lambda_1 + \lambda_2 = 5 + 3.75 = 8.75$. Dĩ nhiên, ta cũng có $E(S) = E(X + Y) = E(X) + E(Y) = \lambda_1 + \lambda_2$.

Xấp xỉ phân phối nhị thức bằng phân phối Poisson



Nội dung

1. Biến ngẫu nhiên và phân phối
2. Biến ngẫu nhiên rời rạc và hàm xác suất
3. Các biến ngẫu nhiên độc lập
4. Kì vọng của biến ngẫu nhiên rời rạc
5. Phương sai và độ lệch chuẩn
6. Một số phân phối xác suất rời rạc điển hình
7. **Phương pháp xấp xỉ kỳ vọng và phân phối bằng mô phỏng**
8. Ví dụ minh họa - Bài toán thu thập phiếu thưởng

Phương pháp xấp xỉ kỳ vọng bằng mô phỏng

Để xấp xỉ kì vọng $E(X)$ của một biến ngẫu nhiên X liên quan đến thí nghiệm T, ta có thể dùng phương pháp thống kê như sau

- Thực hiện lặp lại N lần (độc lập) thí nghiệm T, ghi nhận các giá trị mà X nhận $x_1, x_2, ..., x_N$ (còn gọi là **mẫu dữ liệu** - sample), và tính **trung bình mẫu**

$$\bar{x} = \frac{x_1 + x_2 + ... + x_N}{N}.$$

- Khi N đủ lớn, ta có $\bar{x} \approx E(X)$.
- Việc thực hiện lặp lại nhiều lần thí nghiệm T có thể được **mô phỏng** (simulate) trên máy tính.

Phương pháp xấp xỉ phân phối bằng mô phỏng
Biến ngẫu nhiên rời rạc

Để xấp xỉ hàm xác suất f_X của một biến ngẫu nhiên rời rạc X liên quan đến thí nghiệm T, ta có thể dùng phương pháp thống kê như sau

- Thực hiện lặp lại N lần (độc lập) thí nghiệm T và tính các tần suất p_x của biến cố "X nhận giá trị x".
- Khi N đủ lớn, ta có $p_x \approx P(X = x) = f_X(x)$.
- Việc thực hiện lặp lại nhiều lần thí nghiệm T có thể được **mô phỏng** (simulate) trên máy tính.

Nội dung

1. Biến ngẫu nhiên và phân phối
2. Biến ngẫu nhiên rời rạc và hàm xác suất
3. Các biến ngẫu nhiên độc lập
4. Kì vọng của biến ngẫu nhiên rời rạc
5. Phương sai và độ lệch chuẩn
6. Một số phân phối xác suất rời rạc điển hình
7. Phương pháp xấp xỉ kỳ vọng và phân phối bằng mô phỏng
8. **Ví dụ minh họa - Bài toán thu thập phiếu thưởng**

Bài toán thu thập phiếu thưởng

Bài toán thu thập phiếu thưởng (coupon collector's problem). Một cửa hàng phát hành n loại phiếu thưởng khác nhau. Mỗi lần khách mua hàng sẽ được cửa hàng phát một tờ phiếu thưởng ngẫu nhiên trong n loại. Khi khách thu thập được đủ n loại phiếu thưởng thì được cửa hàng tặng quà.

Hỏi: khách cần mua hàng bao nhiêu lần từ cửa hàng để được nhận quà?

Bài toán thu thập phiếu thưởng
Tính toán chính xác

Gọi X là số lần khách cần mua hàng để được nhận quà, tức là số lần cần mua hàng từ đầu cho đến khi thu thập vừa đủ n loại phiếu thưởng.
Gọi X_i là số lần cần mua hàng từ lúc vừa đã có $i-1$ loại phiếu thưởng cho đến khi thu thập thêm được một loại mới để có vừa đúng i loại phiếu thưởng $(i = 1, 2, ..., n)$.
Ta có $X = \sum_{i=1}^{n} X_i$ và X_i có phân phối hình học với tham số

$$p_i = \frac{n-(i-1)}{n} = \frac{n-i+1}{n} \ (i = 1, 2, ..., n).$$

Từ đó ta có

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \frac{n}{n-i+1} = n\sum_{i=1}^{n} \frac{1}{i} = nH_n,$$

với $H_n = \sum_{i=1}^{n} \frac{1}{i}$ được gọi là **số điều hòa** (harmonic number) thứ n.

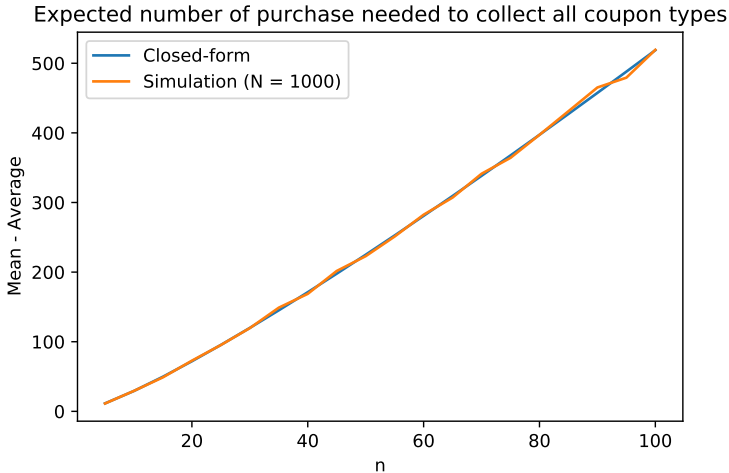
Bài toán thu thập phiếu thưởng - Mô phỏng

```
def num_buy_to_win(n):
    coupons = []
    while len(set(coupons)) < n:
        coupons.append(random.randint(1, n))
    return len(coupons)

def average(n, N, X):
    m = sum(X(n) for _ in range (N))
    return m/N

average(10, 1000, num_buy_to_win)
#29.175
```

Bài toán thu thập phiếu thưởng - Kết quả



Tài liệu tham khảo

Chapter 3-5. Morris H. DeGroot, Mark J. Schervish. *Probability and Statistics*. Addison-Wesley, 2012.

Chapter 3. H. Pishro-Nik. *"Introduction to probability, statistics, and random processes"*, available at `https://www.probabilitycourse.com`. Kappa Research LLC, 2014.